1. INTRODUCTION

Functional Magnetic Resonance Imaging (fMRI) is a rapidly developing tool that enables cognitive psychologists and neuroscientists to study the human brain *in action.* During an fMRI experiment, a subject performs a carefully choreographed sequence of behavioral tasks while Magnetic Resonance (MR) images of the subject's brain are acquired at regular intervals. The tasks are designed to exercise specific motor, sensory, or cognitive processes, and the measured MR signal contains information about the nature and location of the neural activity that results when those processes are engaged. Psychologists hope to use fMRI data to build and test theoretical models of human cognition, but to do this, they must solve a quintessentially statistical problem. In this paper, we develop a new statistical model and new inferential methods for fMRI data.

An fMRI experimental design specifies the tasks each subject is to perform along with their timing and duration. Each distinct task defines one experimental condition, which is usually replicated several times during the experiment. Each contiguous block of time in which the subject performs a specified task is called an *epoch* within the associated condition. In "block trial" designs, the subject repeatedly performs the task during each epoch; in "single trial" designs, the subject performs the task once per epoch.

The choice of tasks is critical to the interpretation of fMRI experimental results because inferences from the data are commonly based on "subtractive" logic. The tasks are designed to differ in the specific components of processing that they (are believed to) require. An observed difference in activation between two conditions is then attributed to the components of processing that differ.

An fMRI experiment yields a sequence of three-dimensional images of the subject's brain. Each image consists of measurements of the MR signal over a grid of small, regular volume elements, called *voxels.* The MR signal for a voxel in a single image is related to the transverse magnetization of a particular nuclear spin, usually hydrogen, averaged over the voxel and a small time interval. However, in *functional* MRI, we are not interested in the images *per se* but rather in small, task-related changes in the measured MR signal over time caused by localized neural activity. The sequence of MR images yields a time series of measurements for each voxel from which we can identify and characterize the brain's response to task-related activity. Voxels typically range from 3 to 50 cubic mm with corresponding images containing from 400,000 to 100,000 voxels. The time to acquire an entire image typically ranges from less than 1 second to over 8 seconds, depending on the acquisition scheme. There is, however, a trade-off between spatial and temporal resolution: images with large voxels can be acquired more quickly than images with small voxels.

The statistical problem in fMRI is to identify and characterize the task-related signal changes in a way that helps scientists test their models and predictions about how the brain works. Current methods for analyzing fMRI data attack what we call the *localization problem*: How can data from an fMRI experiment be used to identify the parts of the brain that activate during performance of a given brain function? Localization in fMRI is a problem of classification, where each voxel is classified as *active* or *inactive* with respect to a comparison between two conditions. The assumption is that in an active voxel there is greater neural activity in response to one condition than to the other and hence that neurons within that volume are involved in the processing distinct to the one condition. Thus, in a well designed fMRI experiment with well chosen conditions it is possible to localize very specific brain functions.

The unit of analysis for most current fMRI methods is the voxel time series; that is, for computational and other reasons, different voxels are treated as independent. The majority of methods uses statistical hypothesis tests to identify "significant" task-related signal changes. For example, the commonly used two sample t-test classifies as active those voxels for which the mean MR signal during the epochs of one condition is significantly different than the mean MR signal during the epochs of another condition. Other frequently used tests include the Kolmogorov-Smirnov test (Aguirre, Zarahn, and D'Esposito 1998), the split sample t-test (Friston, Frith, and Frackowiak 1994), tests of non-zero correlation between the time-series and a fixed reference curve (Bandettini, Jesmanowicz, Wong, and Hyde 1993), t and F tests in ANOVA (Cohen et al. 1994) and more general linear models (Worsley and Friston 1995), and for periodic designs, tests for large power in a frequency band (Weisskoff et al. 1993; Friston et al. 1994). Non-testing classification methods, *e.g.*, variants on cluster analysis, have also been considered (Poline and Mazoyer 1994; Weaver, Saykin, Burr, Riordan, and Maerlender 1994).

It is tempting to begin an assault on the statistical problems of fMRI by trying to improve the testing procedures used for classification, but there are two limitations of the localization paradigm that need to be addressed. First, the data obtained from an fMRI experiment are a realization of a very complex spatio-temporal process with many sources of variation, both biological and technological. The noise in the data is complicated (see section 2), and the task-related signal changes are generated by a response that varies over the brain and has both non-linear and non-additive features. The problem is further complicated by the highly irregular tissue boundaries in the brain and the presence of confounding factors such as large blood vessels.

While the various hypothesis tests and classification procedures often give reasonable results and can reliably detect large signal changes, the simplistic assumptions underlying these methods introduce non-trivial inefficiency into the inferences. This has evoked two responses in the fMRI literature: (1) the development of better pre-processing algorithms to correct the data for "artifacts" (e.g., trends) prior to statistical analysis (Eddy, Fitzgerald, Genovese, Mockus, and Noll 1996) and (2) the search for new classification procedures that address some of the problems of simpler methods. The latter include adjusting the voxelwise tests to improve the shape of identified active regions (Forman et al. 1995), using the distribution of the extremes of random fields to select a global threshold for voxelwise tests (Worsley 1995; Worsley, Evans, Marrett, and Neelin to appear), and applying randomization tests to protect against violations of the assumptions while maintaining a nearly specified type I error for testing the omnibus hypothesis over the brain (Holmes, Blair, Watson, and Ford 1996). In addition, Lange and Zeger (1997) develop a spectral-domain model for the voxel time series that allows variation in the response and thus adapts better than the simple tests. Inferences under their model are based on tests using parameter estimates. (See also the discussions to their paper by Eddy and Genovese.)

The second and more critical limitation of localization methods, even with the above improvements, is that they address primarily a single question: where did the activity occur? While this is an important step in the analysis, many recent fMRI experiments are motivated by scientific questions that relate to more complex relationships in the pattern of responses than localization can reveal, including the impact on the responses of changing task demands, dissociations among regions in the temporal pattern of responses, and functional connectivity among distributed components in a processing network. In particular, such questions arise often when cognitive psychologists test and refine their cognitive theories with fMRI data. Consider two specific examples. First, in the example experiment to be discussed below the relevant cognitive theory predicts that neural activity will increase monotonically across four selected conditions. A hypothesis test for classifying monotonicity could be constructed by combining the results of one-sided, pairwise tests between successive conditions (see the pairwise t-map in Figure 4c), but since equality does not preclude monotonicity, the null hypothesis is composite. The error rates of the combined test are also difficult to compute, providing an unsatisfying assessment of uncertainty. Second, a common practice in the fMRI literature is to compare conditions by counts of active voxels, as classified by a t-map like the one just discussed. While these counts estimate the extent of activation, they provide no useful measure of associated uncertainty with which to provide a baseline against chance fluctuations. Nor is it feasible to construct simple testing procedures for the difference in the counts because the null hypotheses are highly composite and the natural test statistics have intractable null distributions in general.

1.1. An Example Experiment

The brain has a set of general mechanisms, known as *working memory* (Baddeley 1986), for maintaining information during processing. Working memory resources are limited, and performance can suffer when the demands of a task exceed the supply. For example, some sentences are more difficult to understand than others because they require more working memory resources: in parsing complex sentences, any nested clauses, modifiers, or unresolved ambiguities must be held in memory until they can be assigned a role in the meaning of the sentence. Research on language is thus intimately tied to theories of working memory.

Cognitive theories provide an abstract representation of the processes that underlie task performance and make specific, quantitative predictions, *e.g.,* what processes will be used and when, what distribution of responses will result, and how long subjects will take to complete processing. A current cognitive theory of working memory (Just and Carpenter 1992; Just et al. 1996) posits a hierarchy of resource pools specialized to particular types of information. In the context of sentence processing, this theory predicts that sentences with certain syntactic constructs will require monotonically increasing amounts of working memory and that these working-memory demands will manifest themselves in the brain as monotonically increasing activation.

Here, we consider an fMRI experiment designed to test this theory and in particular to study how working memory utilization changes with sentence difficulty. During the experiment, the subject reads a sequence of visually-presented sentences and responds to a question about each sentence by pushing an appropriate button. This experiment, part of a study described in Just et al. (1996), manipulates the difficulty of the sentences presented to the subject. The experimental design specifies six task conditions arranged in thirty-eight task epochs, as illustrated in Figure

1: (R) Simple rest, which serves as a buffer between tasks and as a baseline control; (F) Maintaining visual fixation on a marked point in the center of the visual field; (Tr) Reading strings of consonants—trivial sentence processing with no semantic content; and three conditions (T_1, T_2, T_3) that involve reading and comprehending increasingly difficult sentences. The sentences at Task Levels 1–3 are distinguished by different syntactic and semantic structures that increase the cognitive load required to understand them. Note that by the subtractive logic of the experiment, Tr and F conditions serve as controls for the T_i's. During each epoch, the subject processes several sentences of the given type in succession; the order of the epochs is randomized. The data are described in Section 4.

2. MODELING FMRI DATA

Let $Y(t)$ be the observed MR signal at time t from a specific voxel, where $t = 0, \Delta, \ldots, (T-1)\Delta$ for $\Delta > 0$. Our voxelwise model decomposes this time series into four distinct components

$$Y(t) = \mu + d(t) + a(t\,;\ \mu, \boldsymbol{\gamma}, \boldsymbol{\theta}) + \sigma\,\epsilon(t), \tag{1}$$

where μ, $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$, and the function $d(\cdot)$ are model parameters and ϵ is a parameterized noise process with mean 0 and variance 1. Once the distribution of ϵ is specified, this equation determines the likelihood for the model. The specification of deeper levels in the hierarchy is given in Section 3. The four additive components in equation (1) will be called the baseline signal, drift profile, activation profile, and noise, respectively. Below we clarify the role and parameterization of each component. Throughout, we consider two different strategies for fitting the model: maximization, in which we compute estimates and approximate standard errors by posterior optimization, and sampling in which we obtain draws from the posterior distribution via Markov chain Monte Carlo (MCMC) techniques.

2.1. Baseline Signal

The real-valued parameter μ in equation (1) represents the magnitude of the baseline signal at the given voxel, defined as the mean signal over time in the absence of activation and noise. The baseline signal can vary by an order of magnitude across the imaged volume. While some of this variation reflects differences in nuclear density across the brain tissue, much of it arises from other sources, including differential position and orientation with respect to the receiver electronics, inhomogeneities in the receiver electronics, and local magnetic anomalies in the tissue. Nonetheless, the baseline μ is usually well-determined from the data.

2.2. Drift Profile

The measured MR signal at a voxel tends to drift over the course of an fMRI experiment, and the magnitude of these changes often far exceeds both the ambient noise level and the amplitude of the task-related signal change. The drift profile $d(t)$ in equation (1) represents signal drift as a function of time. Our empirical study of many fMRI data sets suggests that the drift profile exhibits diverse shapes and is often highly nonlinear and heterogeneous over time. It tends to be smooth but has occasional sharp changes which may be related to subject movement. The shape and magnitude of the drift can vary drastically across voxels.

To capture these diverse shapes while avoiding confounding drift with task-related signal changes, we model $d(t)$ as a spline of some degree D (deBoor 1978). For maximization, $d(t)$ belongs to a fixed spline space with a large number (e.g., T) of knots and is regularized similarly to a smoothing spline (Hastie and Tibshirani 1990) (See section 3.) This strategy is computationally efficient and is better able to capture temporally heterogeneous changes than a polynomial basis. For sampling, the number and positions of knots are model parameters with the number of knots kept small (typically 1–5). While more computationally intensive than the fixed-knot approach, this strategy adapts better to the diversity of profile shapes, particularly in its ability to fit both simple and complex forms through shrinkage across spline spaces. In both cases, we constrain $d(t)$ to be orthogonal to the baseline with respect to the empirical inner product, $\langle d \mid 1 \rangle = \sum_{t=0}^{T-1} d(t\Delta) = 0$.

2.3. Activation Profile

Most current fMRI studies rely on the Blood Oxygenation Level Dependent (BOLD) effect (Ogawa et al. 1992) to identify the signs of neural activity. Neural activity initiates a localized inflow of oxygenated blood to the active area, a *hemodynamic response*. Because the magnetic properties of oxygenated and de-oxygenated blood differ (Thulborn, Waterton, Matthews, and Radda 1982), this response is detectable in the MR signal. Figure 2 shows the basic shape of the hemodynamic signal change as a function of time for a response to a single task epoch. It also describes our current parameterization for the response curve, which attempts to capture these basic features while allowing for variation.

We model the response to an isolated epoch of task performance beginning at time t_0 as proportional to the "bell" function $b(t - t_0; \boldsymbol{\theta})$. The parameter vector $\boldsymbol{\theta}$ determines the shape (but not the amplitude) of the response to an isolated epoch of task performance. These shape parameters can vary from voxel to voxel, allowing the model to fit changes in the underlying response across the brain. The vector $\boldsymbol{\theta}$ contains from 2–8 components as illustrated in Figure 2.

Let 0 and t_1 be the times at which task performance begins and ends. These times are given by the design, and we take them as known. The bell function takes the form

$$b(t;\, \boldsymbol{\theta}, t_1) = b_{\text{attack}}(t;\, \boldsymbol{\theta}_{ABC}) \cdot [b_{\text{decay}}(t;\, \boldsymbol{\theta}_{DEF}) - b_{\text{dip}}(t;\, \boldsymbol{\theta}_{GH})] \tag{2}$$

$$b_{\text{attack}}(t;\, \boldsymbol{\theta}_{ABC}) = R_a((t - \theta_A)/\theta_B) \tag{3}$$

$$b_{\text{decay}}(t;\, \boldsymbol{\theta}_{DEF}) = R_d((t - t_1 - \theta_D)/\theta_E) \tag{4}$$

$$b_{\text{dip}}(t;\, \boldsymbol{\theta}_{GH}) = \theta_G \cdot b(t - t_1 - \theta_D; (0, \theta_E/2 - \theta_H, 0, \theta_E/2 - \theta_H, 0, \ldots, 0), \theta_E/2 - \theta_H). \tag{5}$$

where R_a (R_d) is a cubic spline, with an internal knot at $1/2 + \theta_C$ ($1/2 + \theta_F$), that increases (decreases) smoothly from 0 to 1 (1 to 0) over [0,1] and has zero derivative outside [0,1]. The b_{dip} term is excluded when no dip is fit. This parameterization is flexible enough to handle a wide variety of designs, and it enables the fit to capture subtle changes in the response shape.

The activation profile, $a(t)$ in equation (1), combines the responses from every task epoch in the experiment, either additively or non-additively. In the additive regime, $a(t;\, \mu, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \mu \sum_k \gamma_{c_k} b(t - t_k;\, \boldsymbol{\theta})$, where the sum is over epochs, c_k is the condition associated with epoch k, and t_k is the start time of epoch k. The γ_c's, called the *responsiveness* parameters, specify the amplitude of the signal change associated with each experimental condition as a proportion of the baseline signal. By default, we constrain $\gamma_c \geq 0$ for every condition c.

2.4. Noise Distribution

The noise in fMRI data is not simple. Important features of the noise distribution include subject movement, signal drifts, spatial correlations, outliers, and physiological effects. The noise distribution is also sensitive to the scheme used to acquire images. For the results in this paper, we account for drift but use only a simple white-noise process to model $\epsilon(t)$. Our software is capable of substituting other noise families (e.g., heavy-tailed distributions or specific ARMA processes) in the model, but we have not thus far explored ways to select the noise family as part of the fit. We do however pre-process the data with FIASCO (Functional Image Analysis Software, Computational *Olio*) (Eddy et al. 1996) to adjust for some of the more complex noise sources, such as subject movement (Eddy, Fitzgerald, and Noll 1996). While not as good as including these sources of variation directly into the model, these procedures are quite effective in practice.

3. PRIOR INFORMATION

The use of prior information is a critical aspect of inference, particularly in complex or high-dimensional problems, because it enforces substantive constraints and restricts the parameter space to a reasonable form. Our model for fMRI data is over-parameterized, and we need effective prior information to help distinguish the components. Fortunately, each component of our model has itself been the object of research in the MR literature, and our experience working with these data has yielded further insights. The structure of the fMRI data depends critically on the scanner, experimental paradigm, and image acquisition scheme; hence, fitting the model requires some careful work for each site and even each data set. In this section, we describe the available prior information for each model component and illustrate how we use it in one such case.

3.1. Baseline Signal

Uncertainty about the baseline signal μ arises primarily from five sources that relate to varying tissue composition across the brain and spatial effects during scanning and reconstruction. However, because μ is usually well-determined from the data, inferences about μ are not very sensitive to the choice of prior. For simplicity, we use a scaled t_1 distribution centered on a fixed value μ_0, which provides a conservative assessment of our prior uncertainty. The value of μ_0 can be set separately for each voxel with scout images obtained prior to the experiment. This is ideal, but by default we set μ_0 to a typical large signal intensity in the brain, which, because of the arbitrary scaling of the signal in producing the data, must be chosen for each scanner.

3.2. Drift Profile

The prior for d must give weight to the observed properties—general smoothness with several potential change points, some of which may be sharp—while discouraging spurious structure (e.g., oscillatory behavior) that may be confounded with activation. Let K be the number of (internal) knots for the spline $d(t)$ in the interval $[0, (T-1)\Delta]$, and let $\boldsymbol{\kappa}$ with $0 < \kappa_1 < \cdots < \kappa_K < (T-1)\Delta$ be the knot positions.

Our prior for $d(t)$ depends on our fitting strategy. For maximization, we take both K and $\boldsymbol{\kappa}$ to be fixed *a priori*, and we fit the drift with an analogue of a smoothing spline. We typically take K to be proportional to T, and unless there is reason to expect unusual effects (e.g., large subject movements) at the beginning or end of the study, we take the knots to be equally spaced. Given K and $\boldsymbol{\kappa}$, the drift profile d lies in a specific vector space. The prior $\pi_{\text{drift}}(d \mid K, \boldsymbol{\kappa}, \lambda)$ is proportional to $e^{-\frac{1}{2\lambda}Q(d)}$ where $Q(d) = \rho_{nc} \int_0^1 d^2(t)\,dt + \int_0^1 |d''|^2(t)\,dt$and where $\rho_{nc} > 0$ is a fixed constant that determines the relative penalty ascribed to norm and curvature of d. Here, $\lambda > 0$ mediates the overall level of smoothness in the profile, with smoothness increasing as λ gets smaller. We choose the λ that yields a targeted effective degrees of freedom, defined as the trace of the smoothing matrix (Hastie and Tibshirani 1990) because this form is most efficient to compute. This requires an initial noise-level estimate and extra iterative step in the optimization. We can include λ explicitly in the posterior maximization, but this has not proven to be worth the computational cost.

For sampling, K, $\boldsymbol{\kappa}$, and λ are model parameters. Because the knot positions can be chosen adaptively, the number of knots is usually smaller in this case. The prior $\pi_{\text{drift}}(d)$ is of the form $\pi_{\text{drift}}(d \mid K, \boldsymbol{\kappa}, \lambda) \cdot \pi(\boldsymbol{\kappa} \mid K, \lambda) \cdot \pi(\lambda \mid K) \cdot \pi(K)$, where the first term is given above. The conditional prior for the knot positions is derived from a Dirichlet$(\alpha_1, \ldots, \alpha_{K+1})$ distribution for the separations $\kappa_i - \kappa_{i-1}$ (with κ_0 and κ_{K+1} being fixed endpoints) where, by default, $\alpha_i = 2$. The conditional prior for λ is an exponential$(1/\lambda_0)$, where λ_0 is chosen *a priori* as above. The prior for K is a Poisson, truncated at an upper bound K_{max}. This formulation causes π_{drift} to put more mass on profiles with less structure, discouraging spurious features.

3.3. Responsiveness

We choose our prior for $\boldsymbol{\gamma}$, π_{resp}, to match the available information. For example, outside of large blood vessels, responsiveness rarely exceeds 5% for current imaging configurations and is more typically on the order of 2% for a strong signal. We take each $\gamma_c \geq 0$ because the BOLD mechanism leads to a non-negative apparent signal change (though this constraint can be lifted). The prior π_{resp} controls the density of each γ_c conditional on $\gamma_c > 0$ and the determination of which γ_c's are strictly greater than 0. First, conditional on $\gamma_c > 0$, we give γ_c a suitable gamma density by default (typical parameters 1 and 50). Second, we specify a discrete distribution on subsets of conditions, with the full set excluded. For a given subset $A \subset \{1, \ldots, C\}$, then only γ_c with $c \in A$ are allowed to be non-zero; the prior for the others (given this subset) is a point mass at 0. The inclusion of such sub-models is critical because the posterior needs to account for the substantial uncertainty as to whether or not a given voxel shows any response to the specified tasks.

3.4. Shape

The form of the response shape described in section 2 is based on an empirical characterization of the hemodynamic response function. In practice, the manifestation of the response in the data is influenced by the scanner strength, the voxel size, the speed and structure of acquisition, and the nature of the task being performed. For example, in "single trial" studies, where the experiment is designed to detect the response of each individual task performance, responses are weaker and faster with shorter attacks and decays. In "block trials" on the other hand, where the experiment is designed to detect the accumulated response of repeated similar trials, the response function is smoother with longer attacks and decays. By default, we use independent gamma distributions as priors on the attack, decay, both lag parameters, and the undershoot height. The fixed hyperparameters are chosen from one of several templates depending on features of the design and acquisition. There is little prior information to constrain the rise, fall, and skew parameters, but since they are naturally bounded, we use uniform distributions over their range.

3.5. Noise Parameters

For the results in this paper, we use a white-noise model for the noise (excepting drift and motion). We put a gamma prior on the noise precision $1/\sigma^2$ with fixed hyperparameters (1.6 and 200 by default) to approximate the typical measured signal-to-noise level for the scanner we used (GE 1.5T with the Echo Planar Imaging sequence).

4. MAKING INFERENCES FROM FMRI DATA

To illustrate the usefulness of our model in addressing scientific questions beyond localization, we consider the working memory experiment from section 1.1. Recall that the primary goal of that experiment is to test whether activation increases monotonically with sentence difficulty from Tr to T_1 through T_3. We analyze the data from a single subject (#1078), provided by Drs. Carpenter and Just, to fit our model with both the maximization and sampling strategies. Our focus here is not on the results *per se*; rather, we wish to illustrate how our model can be used to directly address scientific questions for which current methods offer only limited insights.

A three-dimensional image consisting of seven slices (each 5mm thick separated by a 1mm gap) was acquired every 1.5 seconds during the experiment. The sequence of images was reconstructed and registered (see Section 5), and each slice was clipped to a 64×64 voxel region containing the brain. (Each voxel is 3.125mm $\times$ 3.125mm in the slice plane and 5mm thick.) Figure 3 provides several linked views of the data for a target slice chosen by the investigators to intersect the two principal language areas: Broca's (production) and Wernicke's (comprehension). Note in particular how the time courses at the boundary of the brain (where the mean signal sharply rises) tend to exhibit more pronounced fluctuations and outliers, probably the result of residual motion (much of it during the "Rest" condition). These edge effects are also apparent in the variance map, which in addition shows a brain-like pattern because the intrinsic signal variance increases with signal level. The breakdown of signal by condition reveals some systematic effects of sentence complexity, although the differences are obscured somewhat by other sources of variation. Figure 1 presents a single voxel time course in detail.

In fitting the model, we use cubic splines with knots at every fourth grid point and relative norm-curvature penalty $\rho_{nc} = 0.01$. The smoothing parameter λ (λ_0 during sampling) is set to target 5 effective degrees of freedom. Conditional on the γ_c's being positive, we give them an exponential(50) prior (a 2% change on average, a reasonable effect for a language task on the scanner setup we used), and we compare the null and saturated sub-models with weights 0.9875 and 0.0125 respectively. We use a four-parameter shape configuration (lag-on, attack, lag-off, decay, all in seconds) and fix the gamma hyperparameters at $(2, 4.3)$, $(4, 1)$, $(2, 4.3)$, and $(4, 0.43)$ respectively based on the estimated responses for similar block designs with earlier data. The gamma prior on the noise precision is given hyperparameters $(1.6, 200)$ to match the distribution of noise levels predicted by previous studies on the same scanner.

An examination of the model fit reveals the following basic relationships. The estimated (posterior mean) baselines are stable (median posterior standard deviation approximately 0.35 or about 0.1% of the baseline value inside the brain) and reveal the brain's anatomical structure as expected; Figure 4a shows one slice. The drifts in these data are non-trivial with an effective signal change of approximately 5% baseline on average for voxels in the brain with only 23% of this variation explained by a linear function. The amplitudes of the task-related signal changes $\boldsymbol{\gamma}$ are mostly very small, as would be expected with a few regions of focal activation. Among those voxels with a probability bigger than 0.001 of a non-trivial response in at least one condition, the 99th percentiles of the posterior means of γ's for T_1 through T_3 are approximately 0.027, 0.028, and 0.029 respectively. A few voxels exhibit large γ's (≈ 0.15) in all four T conditions which are likely artifacts from large blood vessels. The median posterior means of the shape parameters for voxels with a non-trivial response are 0.23, 3.01, 0.24, and 7.42 seconds. These posteriors are quite similar across voxels—the inter-quartile ranges of these posterior means are 0.03, 0.46, 0.02, and 1.16 seconds—which is consistent with a common physiological process underlying the response. Moreover, there is no apparent relationship between the estimate responsiveness and shape parameters, either visually or linearly (maximum absolute correlation 0.36 between any pair). The estimated noise levels across voxels mostly fall in two concentrated clusters corresponding to voxels in air and in the brain. The largest estimated noise levels correspond to voxels just outside the brain or on its boundary, where local movement and other artifacts are most pronounced. The brain voxels on the edge experienced the greatest variance reduction under the fit. (See Figure 5 for a map.) Figure 1 shows a model fit for one voxel time-course.

4.1. Localizing Condition Contrasts

In a typical fMRI analysis, interest centers on contrasts among the task-related signal changes for the experimental conditions. For example, with the sentence processing data, Figure 4b shows a map of two-sample t-statistics comparing the T_3 and Tr conditions, after linear detrending of the data and other pre-processing (Eddy et al. 1996). The map is thresholded at ± 4 for classification (an arbitrary value used by the investigators at the time on all their data sets) to identify active voxels for later consideration. With our model, we can make a similar comparison by examining the distribution of the contrast $\gamma_{T_3} - \gamma_{Tr}$. Conventionally, voxelwise statistics are viewed as images to reveal any spatial structure. We can use a variety of statistics to summarize the posterior distribution. For example, the posterior mean of $\gamma_{T_3} - \gamma_{Tr}$ gives point estimates without a measure of uncertainty. Alternatively, the probabilities $\mathrm{P}\{\, \gamma_{T_3} - \gamma_{Tr} > 0 \mid \boldsymbol{Y} \,\}$ and $\mathrm{P}\{\, \gamma_{T_3} - \gamma_{Tr} < 0 \mid \boldsymbol{Y} \,\}$ measure of the strength of evidence in favor of each ordering without a measure of effect size. These are useful for identifying voxels of interest; the means tell us what is happening at each voxel. Large activation is sparse, but for voxels at which the probability of a non-zero γ in some condition is greater than 0.001, the posterior means range from -0.11 to 0.06 with approximately half positive and half negative. Most of the posterior means are very small because the posterior probability of the null submodel Figure 4c shows a map of $\mathrm{P}\{\, \gamma_{T_3} - \gamma_{Tr} > 0 \mid \boldsymbol{Y} \,\}$. It is straightforward to look at other contrasts $\sum_c \alpha_c \gamma_c$ (e.g., the interaction $\gamma_{\mathrm{T}_3} - 2\gamma_{\mathrm{T}_2} + \gamma_{\mathrm{T}_1}$) in the same way.

4.2. Assessing Monotonicity

A natural approach to the monotonicity question is to ask whether the amplitude of the response increases with the difficulty of the task. In other words, is $\gamma_{\mathrm{Tr}} < \gamma_{\mathrm{T}_1} \leq \gamma_{\mathrm{T}_2} \leq \gamma_{\mathrm{T}_3}$ for a given voxel? Figure 4d shows a map of these "monotonicity probabilities" $\mathrm{P}\{\, \gamma_{\mathrm{Tr}} < \gamma_{\mathrm{T}_1} \leq \gamma_{\mathrm{T}_2} \leq \gamma_{\mathrm{T}_3} \mid \boldsymbol{Y} \,\}$ as computed from the data. (An interesting variant, not shown, is $\mathrm{P}\{\, \gamma_{\mathrm{T}_3} \geq \gamma_{\mathrm{T}_2} \geq \gamma_{\mathrm{T}_1} > 0, \gamma_{\mathrm{Tr}} = 0 \mid \boldsymbol{Y} \,\}$

which in principle can distinguish areas recruited specifically for semantic processing.) The picture reveals strong support for monotonicity in the main cluster (lower right) but much less in the smaller cluster at the middle right. There are several possible reasons for this distinction, but we leave the interpretation of this result to scientific argument.

The brain's response is not restricted to the voxelwise coordinate system we impose, so it is possible that restricting our analysis to voxelwise comparisons may obscure important effects. For this reason, it is common in fMRI to consider the spatial extent of activation as well as the voxelwise amplitudes. The typical approach is to compare conditions via the difference in counts of voxels classified active in each condition (relative to a control), without an accompanying measure of uncertainty at the individual subject level. With our model, we can make proper comparisons of the extent of activation subject by subject. For each voxel v, let N_{iv} be the indicator of the event $\{\gamma_{\mathrm{T_i,v}} > \gamma_{\mathrm{Tr,v}}\}$, for $i = 1, 2, 3$, and let $N_i = \sum_v N_{iv}$ for each i. We use the posterior distribution of (N_1, N_2, N_3) to address the monotonicity question. In general, the (N_{1v}, N_{2v}, N_{3v}) all have different distributions, but using the fast Fourier transform, the convolution can be computed efficiently. The joint probability mass function of each vector (N_{1v}, N_{2v}, N_{3v}) is supported on the lattice $\{0, 1\}^3$ and can be extended by zeros to a larger lattice containing $\{0, 1, \dots, V\}^3$, where V is the total number of voxels being considered. Because we assume the voxels to be independent, we run separate simulations for different voxels, but we can still obtain the distribution of (N_1, N_2, N_3) by multiplying the individual Fourier transforms over the larger lattice and inverting the transform. To reduce the computational burden, we take V to be much smaller than the total number of voxels because there is only negligible mass away from 0 for any of the N_{iv}'s for the vast majority of voxels. With these data, only 147 voxels have posterior probability bigger than 0.001 away from $(0, 0, 0)$, and we take $V = 255$. An alternative strategy is to simulate draws from the distribution of (N_1, N_2, N_3) by generating and adding (N_{1v}, N_{2v}, N_{3v})'s. This is computationally efficient for both small and large V but does add some uncertainty to the estimated probability. In our example, $\mathrm{P}\{\, N_3 \geq N_2 \geq N_1 \mid \boldsymbol{Y} \,\} \approx 0.67$, which appears weakly consistent with monotonicity in extent. As mentioned in section 1, there is no good way to address this question by combining voxelwise classifications (e.g., based on t-tests).

These monotonicity results are consistent with the predictions of a resource-usage model. As demands increase, new units are recruited to the task, both within a voxel—thus explaining the observed monotonicity in amplitude—and across voxels—explaining monotonicity in extent. Because the experimental tasks focus on language comprehension, one would expect the strongest effects to arise in Wernicke's area. Indeed, the investigators attribute cluster 1 in Figure 4d to a part of this area. As the brain schematic in Figure 3 illustrates, Wernicke's area cuts across several slices as it follows the folds of cortex. It appears that Wernicke's area folds back up into the target slice at a point somewhat anterior to cluster 1, which may account for cluster 2. This interpretation is reinforced by the monotonic activation at the same site in the slice directly below the target slice (see right panel of Figure 4d). Returning to the target slice, we also see monotonic activation in cluster 3. This is probably the bottom part of Broca's area, a brain area associated with the production of language (see brain schematic in Figure 3). Finding monotonicity here could be the result of sub-vocalization on the part of the subject (increasing with sentence difficulty) or, perhaps more likely, a reflection of shared function across these two intimately tied language areas. Such redundancy is a fairly common feature of brain organization.

4.3. Model Validation

The results of the model fit—the similarity in response shape across voxels, the small spread in noise levels within the brain, the anatomical location of the activation clusters—are consistent with what we would expect to see if the model were true. We conducted exploratory residual analyses to seek out potential sources of model misfit. For example, examining the residuals as a function of time showed few discernable patterns for most voxels. There are occasional large outliers scattered throughout the brain as well as wild signal fluctuations where the smooth drift could not quite keep up. Voxels near the edge of the brain, particularly those showing strong activation, do exhibit some nontrivial residual autocorrelation. We attribute this to a combination of (i) systematic signal fluctuations not fully captured by the smooth drift term, and (ii) variation in the true response magnitudes across epochs of the same condition (our fits here used a single magnitude parameter per condition). It may also suggest incorporating a low order autoregressive component into the noise process. We also examined residual distributions. If the model fits well, the residuals should be approximately iid normals; even if we do not entirely believe the normal model, this serves as a useful diagnostic. Quantile-quantile plots suggest that the normal approximation is reasonable for voxels inside the brain, where the signal is high and the skewness is negligible. For the voxels outside the brain, where the MR signal is very low, the plots reveal a known lack of fit: the noise distribution is somewhat skewed because the reconstructed data are the modulus of a complex normal. As a quantitative summary of this, we performed voxelwise chi-squared goodness-of-fit tests of the residual distribution against the corresponding normal. For the voxels inside the brain, approximately 1% of the null hypotheses were rejected at the 0.01 level, and the voxels with rejected hypotheses showed no obvious pattern. Figure 5 displays the results of some of these residual analyses for the target slice. Comparison with Figure 3 may be informative.

In addition, we used BIC (Schwarz 1978) to compare our model to the model underlying a widely used fMRI analysis. BIC provides a composite index that balances fidelity to the data and complexity of the model in the fit; lower scores are preferred. (For certain priors, BIC approximates the log Bayes factor between two models.) So, our model will only succeed in this assessment if the additional parameters it uses are balanced by a sufficiently large gain in likelihood. The comparison model is based on the following common procedure (e.g., Just et al. 1996): (i) linearly detrend the data; (ii) fit to the detrend residuals a piecewise-constant function that is constant over task epochs that comprise a single condition; and (iii) test the resulting regression parameters via t or F tests. Our model has a lower score in all but 111 voxels scattered throughout the brain. Among those voxels for which the comparison model scored lower, the median difference is 1.56 with only three having a difference bigger than 10. Among the remaining voxels, the median difference is 10.56 in favor of our model. In sum, these analyses suggest that our model fits the data reasonably well.

There are, however, several sources of lack of fit that need to be addressed. First, the response amplitudes can vary among task epochs of the same condition while as fit here, our model uses the same amplitude across such epochs. Second, when the durations of task performance differ drastically among conditions, the response shape can also vary across conditions within voxel. This problem did not arise with this design, but we have seen other cases where it does. Third, the noise distribution appears to be somewhat heavy tailed, and the occasional large outliers can have some effect on the fit. We deal with this presently by simple Windsorization in FIASCO, but it would be desireable to account for this in the noise model itself. Finally, given the complexity of the data-collection process, it is not surprising that the time courses can occasionally exhibit odd changes—jumps, dips, spikes—for which we have no explanation. All but the last of these issues can be handled by straightforward generalizations of our model, which we are currently pursuing.

5. COMPUTATIONAL TECHNIQUES

To analyze data with our methods, we must fit the model to many thousands of voxel time-series with varying structure. There is little opportunity for manual adjustment during the fit, so we require flexible and efficient numerical methods. The raw data from the MR scanner are collected in the Fourier domain and thus require pre-processing to reconstruct the images and correct for several sources of bias and mis-calibration. (Computational cost currently prevents us from integrating these sources of variation directly into the model.) We pre-process the data using FIASCO (Eddy et al. 1996) and take the resulting time-series of images as input to the model fitting and sampling software BRAIN (Bayesian Response Analysis and Inference for Neuroimaging), a public-domain package written by the author. (The BRAIN home page can be accessed from `http://www.stat.cmu.edu/~genovese/`.)

For maximization, we use direct numerical optimization of the log unnormalized posterior for each responsiveness sub-model. The result is a set of estimates and an approximate covariance matrix derived from the inverse observed Fisher information at the mode which is obtained from the computed Hessian. We estimate the posterior probabilities of the submodels by combining the sub-model prior probabilities and approximate Bayes Factors computed with a form of the Laplace approximation (DiCiccio, Kass, Raftery, and Wasserman 1995). We then average over the sub-models using these probabilities as weights. For MCMC sampling, we use a mix of Metropolis and Gibbs steps with a fixed scan order across components (Smith and Roberts 1993; Tierney 1994). Before recording output, we perform an optional pre-scan to adjust the initial Metropolis jumping distributions and a period of burn-in (by default 5000 samples per parameter) to equilibrate the chain. The maximum posterior estimates are used as the starting point.

The optional pre-scan attempts to find effective widths for the Metropolis jumping distributions. Initial jumping widths are derived from the successive conditional variances of the parameters, which are computed via Cholesky factorization of the approximate covariance matrix. The chain is then run for a fixed number of scans during which heuristic performance criteria are recorded, e.g., rejection rate and path length. This run is divided into blocks, and the widths for a block are adjusted using previous blocks to improve performance measured either by deviation of rejection rate from a target (e.g., 50%) or path length (Roberts, Gelman, and Gilks 1997).

With so many chains to be run, efficiency is critical, and we run them as long as is tolerable. Our default is 10,000 samples per component after burn-in, but very high-resolution images necessitate shorter runs. When feasible, we run the chains longer and sub-sample to reduce correlations in the recorded sequence. We have several ways to speed up the computations, including parallelizing the computation, eliminating uninteresting voxels outside the head, and ordering the computation based on the preliminary estimates.

One area that needs development here is convergence diagnosis since multiple chains and graphical monitoring are inconvenient in practice. We currently use only rudimentary measures of chain performance during analysis, but we are working to improve this. As part of a "quality control" effort, we studied the performance of our sampling scheme on a collection of voxel time series from several experiments. Graphical diagnostics, correlations among parameters, and various standard convergence diagnostics (Cowles and Carlin 1995) based on parallel chains with different starting points suggest that the chains are mixing quite well, and equilibrating sufficiently and also that the normal approximation is reasonable in most cases. However, more systematic study is needed.

Default Sampling Scheme. The baseline parameter μ mediates a number of the other parameters and so has a complicated complete conditional. We use a symmetric random walk Metropolis chain for μ, but as part of the move, we multiply all the $\boldsymbol{\gamma}$'s by μ'/μ (candidate to current) to maintain the activation profile. The complete conditionals for the drift and responsiveness parameters can be sampled directly. We first sample $\boldsymbol{\gamma}$ and then the drift profile conditional on $\boldsymbol{\gamma}$ because the non-negativity constraint on the responsiveness complicates its distribution. The conditional distribution for $\boldsymbol{\gamma}$ given everything but the drift is a multi-variate normal truncated to the positive orthant. To sample from this distribution, we draw the components of $\boldsymbol{\gamma}$ one at a time from successive univariate conditional distributions. The Cholesky factorizations of the covariance matrix and its inverse allow us to derive the mean and variance of these conditional distributions iteratively. We then draw from a univariate truncated normal using the inverse distribution function method when the mean is large enough to ensure precision in computing the normal distribution function and a rejection method (based on an exponential approximation to the normal tail) otherwise. The drift profile can then be drawn as a whole from its complete conditional. The shape parameters capture most of the nonlinearity in the model. We choose from among two different types of Metropolis moves for these parameters: (i) a log normal random walk in the parameters individually, and (ii) coupled jumps in related pairs (lag-on and attack, lag-off and decay, etc.). As an example of the latter, we use one move type that keeps lag-on + attack constant while varying their relative size and another that changes the sum while keeping the relative size the same. These diverse moves provide an automatic reparameterization voxel to voxel that reduces the correlation among the parameters and improves mixing. The smoothing hyper-parameter $\log \lambda$ follows a normal random walk, and the noise precision is drawn from its complete conditional.

Model Jumping. For varying the structure of the model in discrete ways, we use the reversible jump framework (Green 1995). In particular, this allows moves across sub-models in the $\boldsymbol{\gamma}$'s and across spline spaces for the drift. Since these components maintain their interpretation in every sub-model, we average over the models to account for uncertainty in the structure. For example, at each sampling iteration, we update $\boldsymbol{\gamma}$ by a Gibbs' step as described above, and then with some probability attempt a sub-model jumping move: inclusion of a zero component and removal of a non-zero component, both equally likely. If a jump is to be made, we select an appropriate component c at random and then generate a candidate for both γ_c and μ. The latter is moved because switching γ_c to or from 0 affects which measurements provide information about the baseline. Let ℓ_1 and ℓ_2 denote the number of data in the zeroed and candidate condition respectively. The simplest move takes (μ, γ) to $((\ell_1 + \ell_2(1+\gamma))/(\ell_1+\ell_2)\mu, 0)$ for removal and $(\mu, 0)$ to $((\ell_1+\ell_2)/(\ell_1+\ell_2(1+z))\mu, z)$, where z is a random responsiveness candidate independent of μ. Mixing can be improved by randomly perturbing the components here. Both of these follow the template given by Green (1995) and satisfies detailed balance.

Similarly, for drift, we take a Gibbs' step every iteration and then randomly decide among moves that changing a knot position, add a knot, and remove a knot. When adding, removing, or moving a knot, the effected knot is selected at random, and the basis is then reformatted to make it easier to update that component. Moving a knot involves randomly perturbing the selected knot within the bounds of its neighbor. The simplest way to add or remove a knot is to change a single coefficient, setting it to zero when removing or drawing it from a distribution independent of the profile when adding. This attains detailed balance but does not mix very well because only a small perturbation to a single component leads to an acceptable change in the profile. We fix this problem by also updating the other components of the profile as part of the move. We parameterize the drift with orthonormalized spline bases that makes such changes more efficient. The dimension-matching requirement is satisfied, mixing is improved, and detailed balance is maintained.

6. DISCUSSION

Our model has several notable advantages over traditional methods for analyzing fMRI data. It can capture complicated drifts and changes in the shape of the hemodynamic response. In contrast to most current methods, the model components are estimated simultaneously rather than serially. The fit to the data can thus be more precise than the implicit fits underlying most classification tests. By producing estimates of meaningful parameters rather than just test statistics, our method facilitates quantitative interpretation of the results. Moreover, the model provides useful measures of uncertainty in contrast to the difficult-to-use error properties of classification methods. Among recent methods, only the model of Lange and Zeger (1997) has similar advantages. In addition, our model handles complex designs, including both block and single trial, and can include important features of the response such as the undershoot dip and non-linear combination of closely spaced responses. Finally, the model accounts for what scientists know about the underlying processes generating the data and can easily incorporate new information as it becomes available.

The inferences we can derive under our model can address questions of localization and thus subsume the traditional classification-based methods. In our opinion, however, the primary contribution of our approach is that it makes accessible to direct analysis a wide range of more general questions as well. In this paper, we tackle one such question, monotonicity. While other methods (e.g., randomization tests) might be fashioned to handle such specific cases, our model can address a much wider range of questions under a single framework, including questions about changes in the response across conditions and across voxels, about temporal patterns in the response that distinguish different types of processing, and about functional connectivity of brain regions. This flexibility allows scientists to directly target the questions they want to answer with accurate and relevant measures of uncertainty that help them evaluate the analysis.

There are several weaknesses in our approach that remain to be addressed. The first is that fitting the model, particularly via MCMC simulation, is very computationally intensive, requiring on the order of a day to analyze a single subject's data. With careful parallelization and improvements in algorithms, however, we expect this problem to become less severe over time. Second, the noise model used for the results in this paper is still rather simple, and we will extend it to deal with the noise complexity, especially physiological variations and spatial dependence. Third, structural independence of the parameters across voxels is likely a simplistic assumption, and we can gain precision by combining across voxels with similar functional properties. Accounting for this spatial structure dynamically is a challenging problem; we discuss our approach below. Finally, inferences under the model depend somewhat on the priors, and it may be informative to systematically evaluate the nature of this dependence. In our experiments, we have found that specific shape of the priors has only a small impact on the results provided the basic range of the parameters is suitably constrained. Our goal has been to include generally accepted information about the basic processes, so the priors we use reflect reasonably uncontroversial constraints. Nonetheless, further model validation remains a priority.

The basic implementation of our model presented in this paper can be extended in several directions. First, we can allow variations in the response amplitude across epochs and within each condition to capture this variation in the data (Genovese et al. 1997). Second, we can allow variations in the response shape across conditions within a voxel; this is important for designs in which the conditions occur on vastly different time scales. Third, we can parameterize the model by an ANOVA decomposition of the response amplitudes to accommodate factorial designs. Finally, rather than treat voxels as independent units, we need to account for spatial structure in the hemodynamic response. The shape of the response function, the amplitude of the response, the impact of physiological variations, and other such features exhibit complicated dependence across voxels. Modeling these relationships increases the precision of inferences because multiple voxels contribute information about features they have in common. It also facilitates addressing complicated questions about the spatial pattern of responses. The task is not to segment the image *per se* but to identify regions with consistent physiologic and functional properties. A particular challenge here is that tissue boundaries in the brain are convoluted and piecemeal, so methods based on Markov random fields (Geman and Geman 1984; Geman and McClure 1987; Johnson, Wong, Hu, and Chen 1991) will be most effective if given enough anatomical prior information, which is difficult. We are currently working to adapt the method described by Johnson, Bowsher, Jaszczak, and Turkington (1995) that incorporates disjoint region descriptors as parameters at a deeper level in the hierarchy.

Another issue is that large variability in the physical geometry of the cerebral cortex across individuals makes it difficult to compare and combine analyses across subjects. The most common method for combining fMRI data across subjects is to map the subjects' brains onto a common coordinate system called the Talairach atlas (Talairach and Tounoux 1988) and then average in this coordinate system. The Talairach atlas was derived from a detailed study of six human brains, and the mapping for a given subject is computed using only a few gross measurements of that subject's brain. This averaging procedure is far from satisfactory, however, because large inter-subject variations remain. One way our model can be used to combine results applies when the question of interest involves only on a function G over the (total) parameter space that does not depend on the explicit coordinate system of the image for a given subject. For example, G might be the integrated response over a pre-specified and anatomically defined region of interest or the indicator that two tasks yield distinct temporal patterns of response. Suppose for simplicity that G depends only on $\boldsymbol{\gamma}$, and that the J subjects in the experiment contribute data $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_J$. If we are willing to assume that these data are drawn i.i.d. from some population distribution, we can estimate the population distribution of G by combining the distributions of G given $\boldsymbol{Y}_i$, e.g., $E(G(\boldsymbol{\gamma})) \approx (1/J)\ \sum_{j=1}^{J} E(G(\boldsymbol{\gamma}) \mid \boldsymbol{Y}_j)$.

Figure Captions

Figure 1. The experimental design for the working memory experiment. The six conditions are labeled R, F, Tr, T1, T2, and T3 as in the text. The corresponding epochs are displayed in alternating bands along the horizontal axis with the time scale. Also plotted is the time course for one voxel with the posterior mean curve superimposed. Note (i) the slight drift at the beginning of the experiment that seems to stabilize after about 3 minutes and (ii) the apparent uncorrected movement artifact in the last epoch.

Figure 2. A typical shape for the hemodynamic response in fMRI data to a single epoch of task performance. The time during which the task is performed is marked, and the curve shows the pattern of signal change that results. This curve is a polynomial bell function, as described in the text. The labels indicate the role of the various shape parameters in our model. The rise, fall, and skew parameters here affect the shape of the corresponding part of the curve.

Figure 3. Linked views of the data for the target slice. All signal values are measured in "local MR units" determined by the scanner used. Note the inset box on the mean image.

Figure 4. Various results derived from the working memory data for the target slice. The solid outline copied on each map encloses the brain to facilitate comparison across panels; this outline is also included on the baseline map of Figure 5. In each panel, the gray-scale refers to a different quantity as described below.

(a) Probabilities $\mathrm{P}\{\, \gamma_{\mathrm{T}_3} - \gamma_{\mathrm{Tr}} > 0 \mid \boldsymbol{Y} \,\}$. The large number of nearly white voxels results from a posterior mass for the corresponding responsiveness parameters concentrated at zero. The gray-scale for this panel shows the probability values.

(b) A traditional t-map thresholded at the arbitrary but often used value of ± 4. The nominal significance levels suggested by theory do not give the expected error rates, most likely because of complexity in the noise distribution that is unaccounted for by the test. The gray-scale for this panel shows the t values.

(c) Monotonicity probabilities $\mathrm{P}\{\, \gamma_{\mathrm{T}_3} \geq \gamma_{\mathrm{T}_2} \geq \gamma_{\mathrm{T}_1} > \gamma_{\mathrm{Tr}} \mid \boldsymbol{Y} \,\}$. The gray-scale for this panel shows the probability values.

(d) Left panel: Monotonicity probabilities for target slice extracted from the right side of the image (left side of the brain) in (c). Cluster identifications are discussed in the text. Right Panel: Corresponding region in the slice adjacent to and below the target slice. See brain schematic in Figure 3.

Figure 5. Linked views of the residuals and fit for the target slice. All signal values are measured in local MR units.